VTT Technical Research Centre of Finland

Data-driven precision medicine ecosystem - PreMed phase 3 report

Lähteenmäki, Jaakko; Fagerström, Richard; Liedes, Hilkka; Pajula, Juha; De Pretis, Francesco; Vuorinen, Anna-Leena; Ylén, Peter

Published: 19/05/2021

Document Version
Publisher's final version

RESEARCH REPORT

VTT-R-00406-21

Data-driven precision medicine ecosystem

PreMed phase 3 report

Authors: Jaakko Lähteenmäki (editor), Richard Fagerström, Hilkka Liedes, Juha Pajula, Francesco De Pretis, Anna-Leena Vuorinen, Peter Ylén

Confidentiality: Public

Report's title

Data-driven precision medicine ecosystem - PreMed phase 3 report

Customer, contact person, address	**Order reference**
Outi Tuovila Business Finland, Porkkalankatu 1	3458/31/2017
Project name	**Project number/Short name**
Data-driven precision medicine ecosystem	115511 / PreMed
Author(s)	**Pages**
Jaakko Lähteenmäki (editor), Richard Fagerström, Hilkka Liedes, Juha Pajula, Francesco De Pretis, Anna-Leena Vuorinen, Peter Ylén	14/
Keywords	**Report identification code**
precision medicine, personalized medicine, secondary use, artificial intelligence, genome technology, biobank, business ecosystem, pharmacogenetics, antithrombotic drugs	VTT-R-00406-21

Summary

This report summarizes PreMed project results and activities in project phase 3 (1.1.2020 - 30.04.2021). The activities have been focused in three main areas: biobank study (data collection, integration and analysis), ecosystem modelling and dissemination (project workshops and other events). The biobank study addresses pharmacogenomics of antithrombotic drugs. The experiences of the data collection (data permit applications, data quality check and integration etc.) were documented to provide input for national and international development of infrastructures for secondary use of health data and biobanking. We carried out data analysis to reveal associations between genotypes and health outcomes in patients using antithrombotic drugs. Associations were found for warfarin, direct oral anticoagulants and ticagrelor. The results are expected to be helpful in identifying individuals with suboptimal response to antithrombotic medication. The project also developed a simulation model for the analysis of different precision medicine ecosystem development scenarios. For example, different public funding scenarios can be compared in terms of their effect to the volume of data-driven projects initiated by the pharmaceutical industry. The project has continued dissemination of project results in two internal project workshops and one public workshop. The project results have so far been published in two journal articles and two conference papers. The data analysis continues in VTT's internal project activities as enabled by the data permits.

Confidentiality	public

Espoo 30.4.2021

Written by	**Reviewed by**	**Accepted by**
Jaakko Lähteenmäki Principal Scientist	Kari Kohtamäki Key Account Manager	Jari Ahola Research Team Leader

VTT's contact address

Jaakko Lähteenmäki, jaakko.lahteenmaki@vtt.fi

Distribution (customer and VTT)

To be published at VTT Research Information Portal.

Contents

1. Introduction

Personal data is increasingly collected and used in the context of digital services. Consequently, the amount of personal data stored in information systems is increasing exponentially. Besides the primary purpose, collected personal data is used for so called secondary purposes, such as for monitoring the quality of provided services, in the development of new services or products and for scientific research. In general, secondary use is permitted under certain conditions by the General Data Protection Regulation (GDPR). Additionally, specific legislation addressing secondary use of health data and biobanking has been implemented in some countries, including Finland[1,2].

VTT initiated the PreMed project in 2017 with the objective of promoting the development of a data-driven precision medicine ecosystem in Finland. In particular, the project aimed at collecting different kinds of companies together with the common objective of exploiting the opportunities provided by health data resources. The project partners represent different roles of the data-driven precision medicine ecosystem as shown in Figure 1.

In the first phase of the project (May 2017-October 2018) VTT carried out a series of interviews for better understanding of the expectations of companies towards exploitation of data and for identifying the bottlenecks and challenges faced today. The results of the interviews and the analysis of the data-driven precision medicine landscape were compiled into the PreMed phase 1 report[3]. In phases 2 and 3 the main project activity was focused in executing a retrospective cohort study on pharmacogenomics (PreMed study). The objective of the study was to use data from biobanks and national registries to assess the relevance of using genetic data in the context of antithrombotic drug therapy. Besides the scientific objectives, the biobank study was expected to provide valuable information and experience on data access processes and related bottlenecks. Additionally, the project has developed a system dynamics model for simulating alternative precision medicine ecosystem development scenarios. The activities related to the planning of the biobank study and development of the simulator were documented in the PreMed phase 2 (November 2018 – December 2019) report[4].

This report focuses on the results of the PreMed phase 3 (January 2020 – April 2021). The report provides an overview of the activities and achieved results. More detailed information can be found in the published articles [1, 2] and conference presentations[5]. Overall project information can be found at the project website[6].

[1] Biobank Act, https://www.finlex.fi/fi/laki/kaannokset/2012/en20120688.pdf
[2] Act on the secondary use of health and social data, https://stm.fi/en/secondary-use-of-health-and-social-data
[3] PreMed phase 1 report, https://cris.vtt.fi/en/publications/data-driven-precision-medicine-ecosystem-stakeholder-needs-and-op
[4] PreMed phase 2 report, https://cris.vtt.fi/en/publications/data-driven-precision-medicine-premed-phase-2-report
[5] A System Dynamics Model of Data-Driven Precision Medicine Ecosystem https://cris.vtt.fi/en/publications/a-system-dynamics-model-of-data-driven-precision-medicine-ecosyst
[6] PreMed website: www.vtt.fi/premed

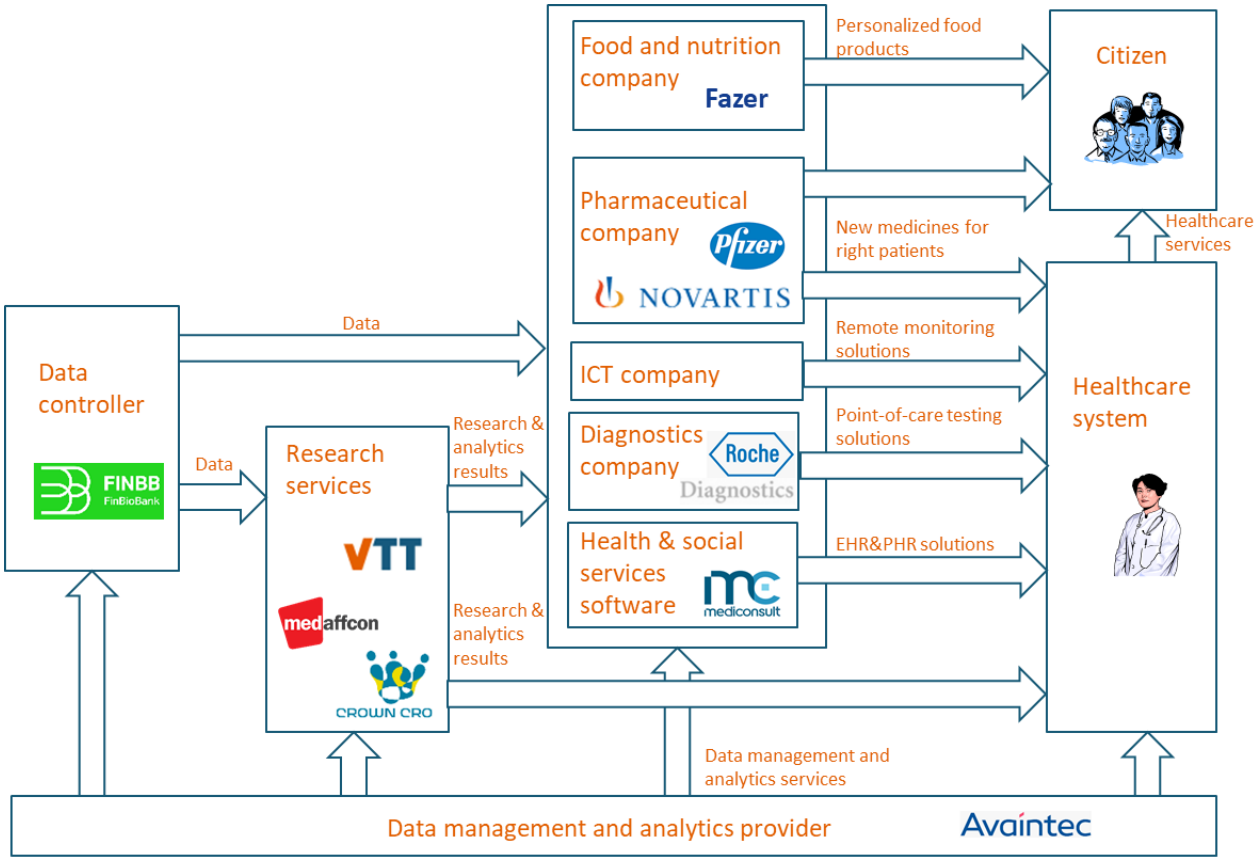

Figure 1. Data-driven precision medicine ecosystem showing roles of PreMed partners.

2. Overview of project activities

Activities carried out during phase 3 of the project are shown in Figure 2. The ecosystem model was further developed during the first half of year 2020 after which the model and simulation results were reported in a conference presentation in July. Most of the phase 3 effort was invested in carrying out the biobank study. The data collection was completed in May followed by the data analysis and reporting phases. Three project workshops were organized during phase 3. The project activities and obtained results are described in more detail in the following sections.

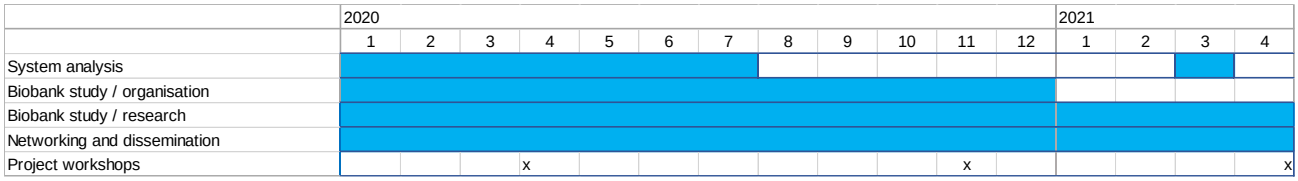

Figure 2. PreMed project activities during phase 3.

3. Biobank study organisation

An important objective of the project was to collect experiences of the process of applying data permits and integrating the data from heterogeneous sources for analysis. The experiences are reported in [2]. The data sources for the study were Auria Biobank, Helsinki Biobank, Finnish Institute of Health and Welfare (THL), THL Biobank, Finnish Prescription Register (Kela) and Finnish Hospital districts and municipalities.

The data collection process and timeline are depicted in Figures 3 and 4. VTT first submitted the application for ethical review and then data permit applications to the biobanks, THL and Kela. After the applications were accepted biobanks started data integration. Each biobank formed a subcohort by linking their genome and patient record data with THL and Kela register data as well as with laboratory data from hospital districts and municipalities. The subcohorts were delivered to VTT in pseudonymized form (without personal identifiers). VTT integrated the full cohort by combining the subcohorts.

The study showed that a large cohort of patient data can be formed by integrating various data resources. Co-operation with all partners was smooth. However, as this was one of the first projects where data resources across several biobanks were combined, the practices for co-operation were not fully established and needed to be defined during the project. The data quality was observed to be good in general [2]. The medication data extracted from hospital data lakes was incomplete (as informed beforehand by the biobanks), but still useful in complementing drug dispensation data of Kela. The full data collection process took 16 months (from January 2019 to May 2020) and required an effort of approximately 5 months from VTT researchers. The data costs, including applications and data extraction of biobank and national register data, were approximately 80 000 €. The long and laborious data collection phase is clearly a challenge for both industrial and academic use of data. Especially, it was affected by the congested data permit service of THL register authority. The data collection could not be fully proceeded before all data permits were granted. This caused idle periods of several months. Improvement will hopefully be achieved via the centralized “one stop shop”

services of Findata, which were not yet available for the PreMed project. The process related to using biobank resources linked with register data will be regulated by the new biobank law currently in preparation[1].

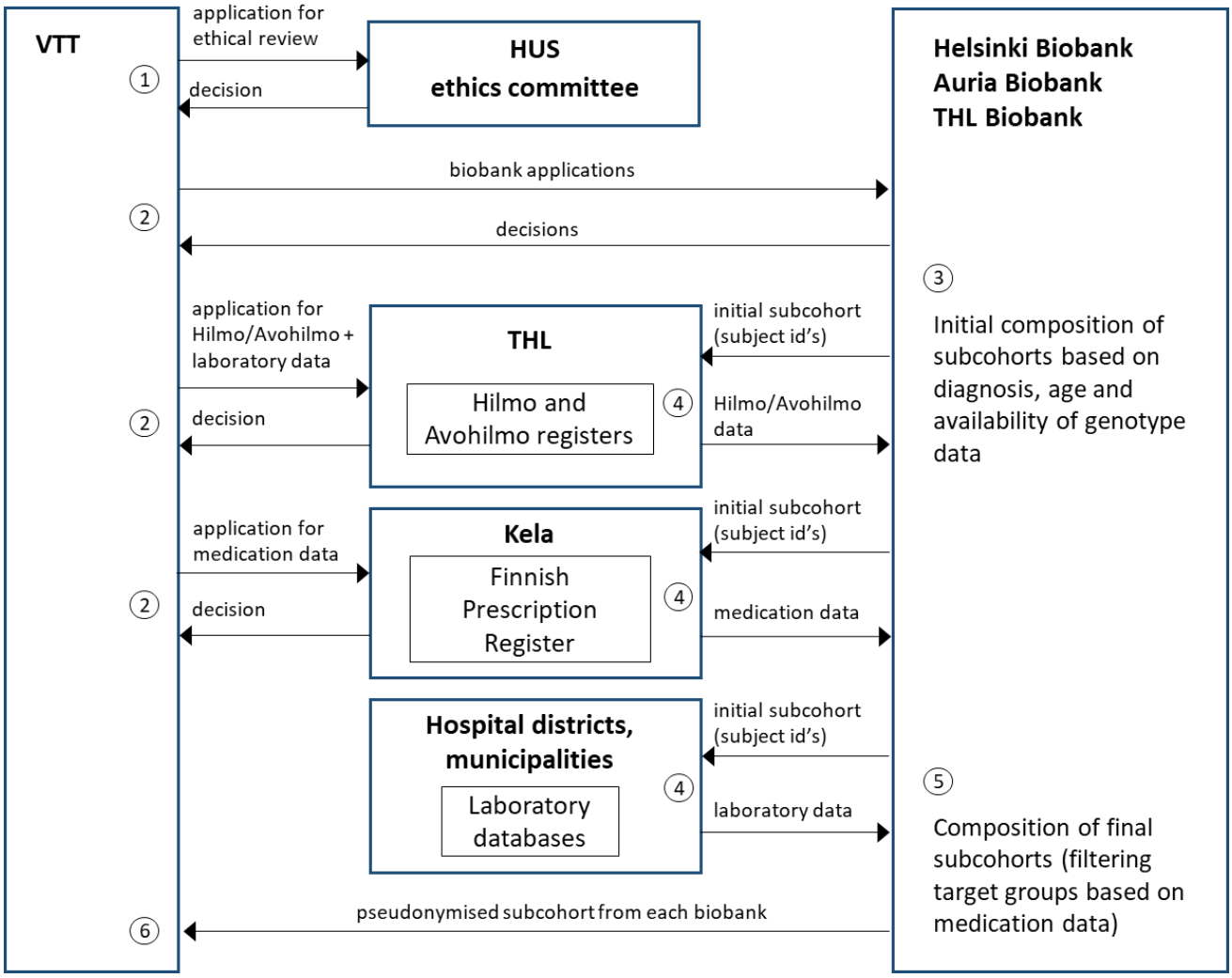

Figure 3. Data collection process [2].

[1] https://stm.fi/hanke?tunnus=STM110:00/2015

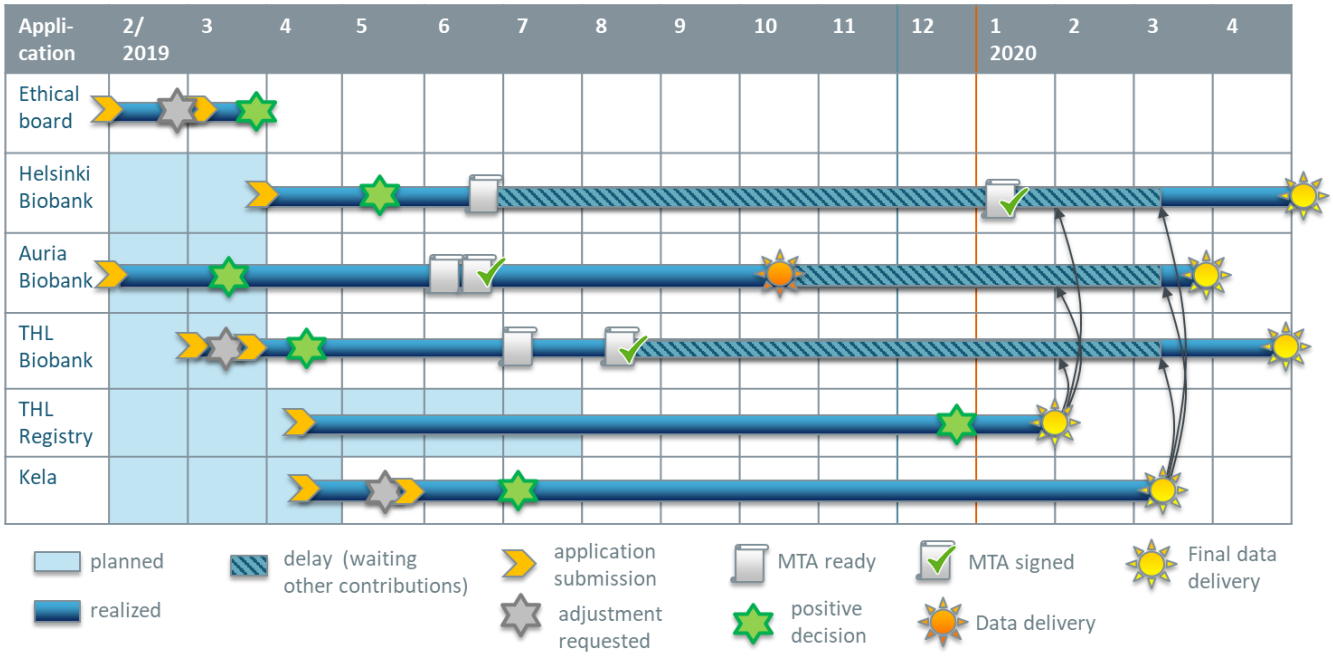

Figure 4. Data collection timeline.

4. Data analysis

In the framework of the PreMed project the following drugs were investigated:

- warfarin
- direct oral anticoagulants, DOACs (dabigatran, rivaroxaban, apixaban)
- ticagrelor

The data analysis was mostly focused in investigating associations between single nucleotide variants (SNV's) and clinical outcomes. Candidate variants to be investigated for each drug were based on earlier studies as defined in the study protocol[1]. We also compared the healthcare costs of different genotype groups in warfarin users to assess possible cost savings achievable by genotype-guided drug therapy.

Overview of the results is presented in the following subsections. Results on warfarin and DOAC pharmacogenetics have been reported in [1] and [3]. Articles on ticagrelor pharmacogenetics and on the impact of pharmacogenetics on warfarin therapy costs are being written.

Additionally, investigation of clopidogrel pharmacogenetics is ongoing as a separate research activity at VTT as enabled by the granted data permits.

4.1 Methods

Investigation of associations between genetic variants and clinical outcomes was based on comparing the incidence of outcome events between patients with different genotypes. Cox proportional hazards model was used in these comparisons and for adjusting of confounders.

[1] PreMed phase 2 report, https://cris.vtt.fi/en/publications/data-driven-precision-medicine-premed-phase-2-report

In the case of warfarin, also parameters derived from international normalized ratio (INR) measurements were used as outcome indicators.

We compared normal responders to the group of sensitive or highly sensitive responders [4] to warfarin. This approach enables the combined effect of the most important variants of the *CYP2C9* and *VKORC1* to be analysed. In the case of DOACs we studied the genetic associations using both genotype and haplotype based groups. The haplotypes were estimated from the genotype data by using the PHASE (v2.1.1) software package [5].

For the investigation of the impact of pharmacogenetics on warfarin therapy we analysed separately healthcare visit costs, medication costs and laboratory costs. Healthcare visit costs were further divided according to the reason for the visit (bleeding or thromboembolic event) and the type of the visit (outpatient visit or hospital stay). The medication costs were calculated based on purchased amount and the drug price. Laboratory costs were calculated from the number of laboratory tests and the nominal price of one laboratory test [6]. The costs of encounters were based on the nominal prices of outpatient visits and hospital days defined separately for primary and secondary care [7].

4.2 Warfarin

Warfarin is still currently the most used anticoagulant in Finland with 110 072 users in 2019[1]. The pharmacogenetics of warfarin has been extensively studied before [8], but not in the Finnish population.

PreMed study cohort included 2508 warfarin users [1]. The results show that the sensitive or highly sensitive patients group had more INR measurements above the normal therapy window (INR=2-3) and that normal responders had more INR measurements below the normal therapy window. This difference, however, did not translate into clinical outcomes as expected: there was no statistically significant difference in the bleeding events between the groups. Instead, thromboembolic events were more common in the sensitive or highly sensitive group. This is expected to be a consequence of the fact that INR test results are used in defining the required dose and, therefore, high INR values lead to lowering the dose. This may have led to too small doses or even pauses in the warfarin treatment, and consequently, increased the risk of thromboembolic events. The detailed results have been published in [1].

4.3 Direct oral anticoagulants

Direct oral anticoagulants are new anticoagulation drugs rapidly becoming more popular due to their favourable pharmacodynamic and pharmacokinetic properties and since they do not normally require continuous therapeutic monitoring. In 2019 there were 19981 dabigatran users, 57619 rivaroxaban users, 62858 apixaban users and 6949 edoxaban users in Finland. Pharmacogenetics of DOACs have been investigated by earlier studies, but only in a few cases associations between genetic variants and clinical outcomes have been found [9]. Most of the earlier studies have investigated the pharmacokinetics of DOACs but not the related clinical outcomes. The PreMed study is to our knowledge the first register-based real world data study addressing the association between genetic variants and clinical outcomes in DOAC users.

The PreMed study cohort included 340 dabigatran users, 999 rivaroxaban users and 467 apixaban users. Edoxaban users were not included in the analysis as data for only 25 patients was available.

[1] Kela – lääkekorvausten saajat ja reseptitiedot (2008-2019), https://www.kela.fi/tilastotietokanta-kelasto_sisallysluettelo#laakkeet

As laboratory measurements are not systematically taken to guide DOAC therapy, it was not possible to use parameters derived from laboratory measurements as outcome indicators. Therefore, only bleeding and thromboembolic events were used as outcome indicators. We found two of the investigated SNV's of the *ABCB1* gene to be associated with thromboembolic events in rivaroxaban users and corresponding results were seen also in the haplotype-based analysis. We also found one of the investigated *ABCB1* SNV's to be associated with bleeding events in apixaban users. The detailed results will be reported in an article submitted for publication [3].

4.4 Ticagrelor

Ticagrelor is the first of new class antiplatelet agents with 8423 users in Finland in 2019[1]. Characteristics of ticagrelor overcome some of the limitations of dual antiplatelet therapy with aspirin and clopidogrel: ticagrelor is rapidly absorbed, it does not require metabolic activation to an active form, it binds rapidly and reversibly to the $P2Y_{12}$ receptor, its pharmacokinetic profile is not significantly affected by age, gender or administration with food, and its pharmacodynamic characteristics are not affected by *CYP2C19* and *ABCB1* SNVs [10].

This study investigated if *CYP3A5*3*, *CYP3A4*22* and *CYP4F2* variants were associated with incidence of bleeding in ticagrelor-treated patients in the PreMed cohort (n=368). We found that carriers of *CYP3A4*22* had increased risk of bleeding compared to *CYP3A4*22* non-carriers. No associations were found between bleeding incidence and *CYP3A5*3* or *CYP4F2* carrier status. Studies with larger populations are needed to confirm this finding. The detailed results will be reported in an article currently being prepared.

4.5 Comparison of warfarin therapy costs between genotypes

We compared costs accrued during the warfarin exposure period between normal and sensitive or highly sensitive patients. The comparison was based on the same set of 2508 patients that was analysed for associations between genotypes and outcome events [1].

According to the results, the medication costs were lower for the group of sensitive and highly sensitive responders due to the fact that their daily warfarin dose was smaller in average. Laboratory costs and costs of healthcare visits related to bleeding events were higher for the sensitive and highly sensitive responders although statistical significance was not achieved. However, the difference between the groups in costs of visits related to bleeding events was close to statistical significance (p=0.064). This finding is in line with the expected elevated risk of over-anticoagulation in sensitive and highly sensitive patients. However, the finding contradicts our earlier analysis where we could not see a significant association between the genotype and bleeding events [1]. The reason seems to be that sensitive or highly sensitive responders have more severe or frequent bleeding episodes requiring more outpatient and hospital visits, while there is no difference in the number of patients experiencing bleeding events. The detailed results will be reported in an article currently being prepared.

[1] Kela – lääkekorvausten saajat ja reseptitiedot (2008-2019), https://www.kela.fi/tilastotietokanta-kelasto_sisallysluettelo#laakkeet

4.6 Usage of pharmacogenetic tests

Pharmacogenetic tests are commercially available and physicians may order them in the context of prescribing drugs with known pharmacogenetic properties[1] [11]. Warfarin is one of the drugs for which such test is available. Based on the laboratory data we found that the pharmacogenetic test (code: B -Farma-D or B -Varfa-D) had been taken only for 7 individuals (0.28%) in the group of 2508 patients.

5. Ecosystem simulation

The data-driven precision medicine ecosystem is expected to lead to remarkable economic benefits through new products and business opportunities. However, considerable public investments are needed to establish the required infrastructure (data/samples, biobank processes), to accelerate the R&D of ecosystem companies and to support related academic research. For decision-makers an important question is how to support the data-driven precision medicine ecosystem growth to achieve maximum benefit for the society. The PreMed project has developed a system dynamics model for simulation of alternative development paths of the ecosystem. Figure 5 shows the high-level ecosystem diagram indicating alternative targets for public investments. The primary outcome measure is the volume of research projects making use of data obtained from biobanks and national registers. The simulator was implemented using the Vensim simulation software[2]

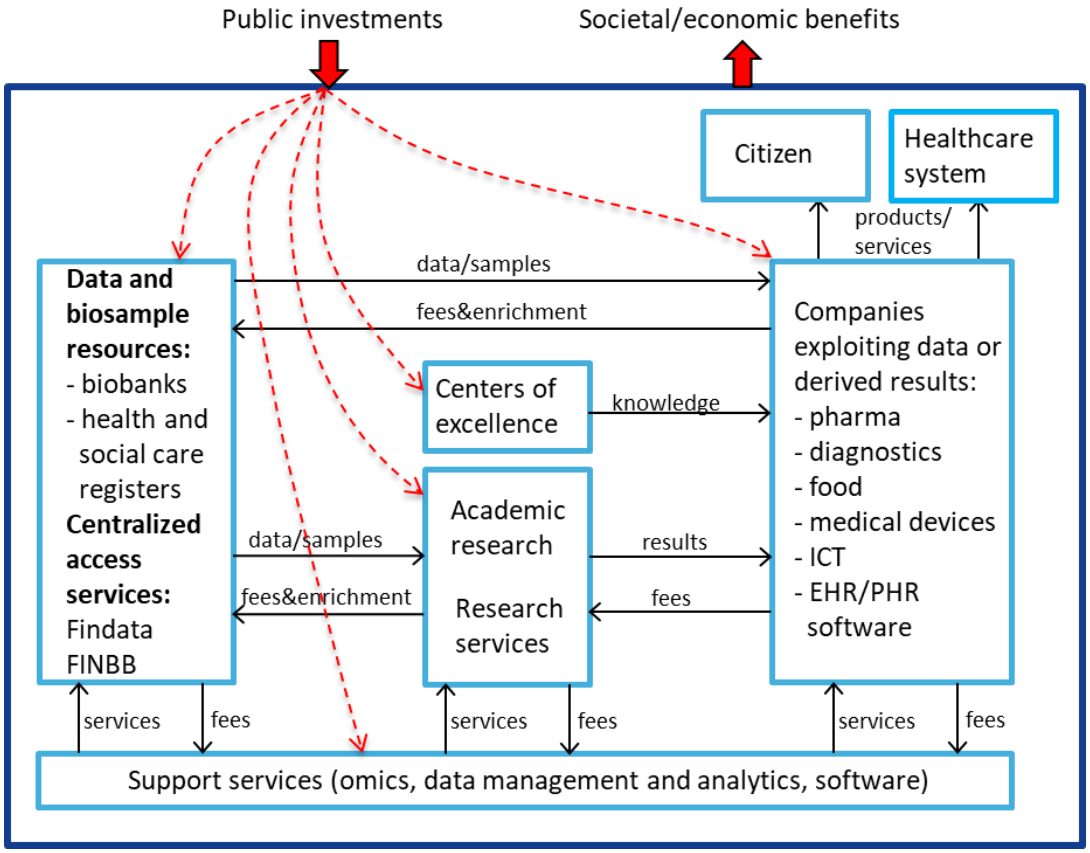

Figure 5. High-level ecosystem model showing main stakeholder groups and dependences affecting the exploitation of health data.

[1] Abomics, https://www.abomics.fi/en/

[2] Vensim simulation software, https://vensim.com/vensim-software/

An example of simulation results is presented in Figure 6. The volume of research projects based on health data resources initiated by pharmaceutical industry in Finland is shown as a function of time. The volume of data-driven research depends on the allocation of resources for R&D by the industry and on the pharma companies' preferences in focusing the resources to RWD (real world data) or other data-driven projects. This choice is assumed to be affected by the availability of health data (including genomic data) and the quality of the related processes (e.g. delivery times and needed resources for data permit applications). In the simulation model the quality of the data access processes is linked with public investments in the centralized services (Findata) and biobanking.

In the **base scenario** of Figure 6 the public investments starting 1.1. 2020 are 5 M€ per year for three years for genotyping and 3 M€ per year for five years for Findata services. In the **increased public investment scenario** the yearly investments for Findata are 6 M€ and the investments continue longer: ten years for genotyping and six years for Findata. In the **increased public and private investment scenario** additionally the yearly growth of the R&D investments of pharmaceutical industry are increased from the base value of 3% to 5%. Realized yearly investments of the pharmaceutical industry in research and in register-based research are shown for reference. Description of the simulator and more simulation results of various scenarios have been reported in the conference presentation materials available at the project web site [12].

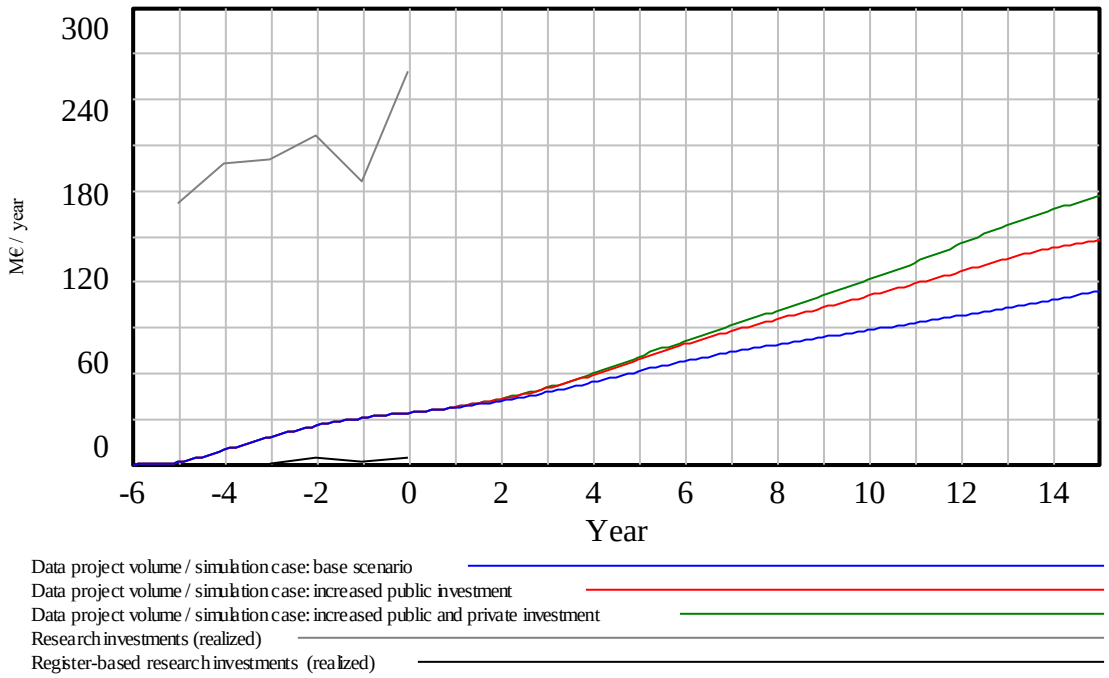

Figure 6. Example of ecosystem simulation results. Volume of Finnish pharmaceutical industry projects exploiting data resources ("data project volume") in three simulation cases. Realized investments of pharmaceutical industry in research and in register-based research are shown as reference[1].

[1] https://www.laaketeollisuus.fi/uutishuone/tilastot.html#inline-3

6. Discussion and conclusions

We have reported results achieved in phase 3 of the PreMed project. The overall objective of the project was to promote the development of the data-driven precision medicine ecosystem in Finland. Under this overall objective, a biobank study with specific needs for data integration from several sources was carried out. Besides the scientific goals, the objective of the biobank study was to collect practical information of the data exploitation processes including related bottlenecks in order to pave the way for data-driven healthcare.

The biobank study demonstrated the capability of the Finnish biobanks together with national register controllers to provide data resources for research purposes. Such projects integrating data from several data sources are not trivial and could not be executed in all countries. Along the way, several challenges were faced as reported. Most importantly, the process of data permit application and data integration was laborious and long, taking 16 months altogether. The PreMed data collection was carried out before Findata services were available. Thus, our experiences and observations provide a bench-marking reference for the Findata services being developed. We have communicated the faced challenges to Findata in order to support the development of new centralized services.

The obtained data sets were in general of high quality and provided a good basis for the data analysis to be carried out in the study. Thanks to the on-going FinnGen project[1] and past research projects of THL, genotype data is largely available in biobanks for research. By linking genotype data with patient record and laboratory data we were able to observe the associations between genetic variants and health outcomes.

The Finnish infrastructure for secondary use of data and biobanking is well-established and serves as a model for other countries, where such infrastructure is still developing. However, more effort and resources are needed to fully exploit the health data resources in Finland. The current processes and practices do not sufficiently support the execution of projects where data needs to be combined from several sources. The Government's bill for a new biobank act[2] defines Findata to be the permit authority also for biobank materials and data. However, the proposed legislation has been largely criticized and, therefore, the content of the law may still be considerably modified. Especially, the academic research community sees challenges in the legislation on secondary use of data. The worries are related especially to increasing costs, sufficiency of Findata's resources and to the limitations (e.g. concerning available computing power) caused by the secure processing environments[3].

There is a need for public-private co-operation in developing an efficient infrastructure for health and social data exploitation. The PreMed project developed a system dynamics model, which can be used for simulation of the evolution of the data-driven ecosystem. The simulator enables different public investment scenarios to be investigated e.g. by comparing their effects on the volume of data-driven projects initiated by the pharmaceutical industry. The input parameters of the model include available statistical data such as the pharmaceutical R&D and real world data project investments as well as the amount of biobank consents.

The simulation model involves several unknown factors. For example, we do not know the effect of data availability on the R&D project initiation and such parameters needed to be estimated. Even with limited input data, we expect that the model is useful in comparison of different alternative development paths. As more information on the data exploitation processes accumulate the model parameters can be defined more realistically. The model was

[1] FinnGen project, https://www.finngen.fi/en
[2] https://stm.fi/hanke?tunnus=STM110:00/2015
[3] Toisiolain vaikutukset tutkimukseen ja data-analytiikan sovelluksiin: Hyteairon analytiikkatyöryhmän selvitys (in Finnish), https://cris.vtt.fi/en/publications/toisiolain-vaikutukset-tutkimukseen-ja-data-analytiikan-sovelluks

constructed from the perspective of pharmaceutical industry, but could be extended to other sectors in further development.

The objective of the biobank study carried out in the project was to investigate the pharmacogenomics of antithrombotic drugs. In the framework of the PreMed project we were able to complete the analyses on 4 anticoagulants (warfarin dabigatran, rivaroxaban, apixaban) and one antiplatelet (ticagrelor). The results on warfarin pharmacogenetics have already been published while another article on dabigatran, rivaroxaban and apixaban has been submitted for publication. Article on Ticagrelor pharmacogenetics and costs aspects of warfarin pharmacogenetics are being written. The analysis on other drugs – especially clopidogrel - is being continued in the framework of VTT's internal research activity. The results obtained so far indicate associations between genotypes and clinical outcomes in warfarin, rivaroxaban, apixaban and ticagrelor users. The results increase our understanding of the impact of genetics on the effects of these drugs and may aid in identifying individuals with suboptimal response to the drug therapy. Further research is warranted on the clinical impact and cost-effects of pharmacogenetic variability in the drug response.

References

[1] Vuorinen A-L, Lehto M, Niemi M, et al. Pharmacogenetics of Anticoagulation and Clinical Events in Warfarin-Treated Patients: A Register-Based Cohort Study with Biobank Data and National Health Registries in Finland. *Clin Epidemiol* 2021; Volume 13: 183–195.

[2] Lähteenmäki J, Vuorinen A-L, Pajula J, et al. Integrating data from multiple Finnish biobanks and national health-care registers for retrospective studies: Practical experiences. *Scand J Public Health* 2021; 140349482110044.

[3] Lähteenmäki J, Vuorinen A-L, Pajula J, et al. Pharmacogenetics of bleeding and thromboembolic events in direct oral anticoagulant users. *Submitt Publ*.

[4] Mega JL, Walker JR, Ruff CT, et al. Genetics and the clinical response to warfarin and edoxaban: findings from the randomised, double-blind ENGAGE AF-TIMI 48 trial. *Lancet* 2015; 385: 2280–2287.

[5] Stephens M, Scheet P. *Accounting for Decay of Linkage Disequilibrium in Haplotype Inference and Missing-Data Imputation*. 2005.

[6] Leminen A, Pyykönen M, Tynkkynen J, et al. Modeling patients' time, travel, and monitoring costs in anticoagulation management: Societal savings achievable with the shift from warfarin to direct oral anticoagulants. *BMC Health Serv Res*; 19. Epub ahead of print 27 November 2019. DOI: 10.1186/s12913-019-4711-z.

[7] Terveyden- ja sosiaalihuollon yksikkökustannukset Suomessa vuonna 2011, https://www.julkari.fi/handle/10024/114683 (accessed 10 March 2021).

[8] Yang J, Chen Y, Li X, et al. Influence of CYP2C9 and VKORC1 genotypes on the risk of hemorrhagic complications in warfarin-treated patients: A systematic review and meta-analysis. *Int J Cardiol* 2013; 168: 4234–4243.

[9] Raymond J, Imbert L, Cousin T, et al. Pharmacogenetics of Direct Oral Anticoagulants: A Systematic Review. *J Pers Med* 2021; 11: 37.

[10] Teng R. Pharmacokinetic, pharmacodynamic and pharmacogenetic profile of the oral antiplatelet agent ticagrelor. *Clinical Pharmacokinetics* 2012; 51: 305–318.

[11] Verbelen M, Weale ME, Lewis CM. Cost-effectiveness of pharmacogenetic-guided treatment: are we there yet? *Pharmacogenomics J* 2017; 17: 395–402.

[12] Ylén P, Lähteenmäki J, Sorasalmi T. A System Dynamics Model of Data-Driven Precision Medicine Ecosystem. In: *2020 International Conference of the System Dynamics Society, SDS 20*, p. 19.